Analysis of a two-class FCFS queueing system with interclass correlation

Herwig Bruneel, Tom Maertens, Bart Steyaert, Dieter Claeys, Dieter Fiems, and Joris Walraevens

Ghent University,
Department of Telecommunications and Information Processing,
SMACS Research Group,
Sint-Pietersnieuwstraat 41,
B-9000 Ghent, Belgium
{hb,tmaerten,bs,dclaeys,df,jw}@telin.UGent.be

Abstract. This paper considers a discrete-time queueing system with one server and two classes of customers. All arriving customers are accommodated in one queue, and are served in a First-Come-First-Served order, regardless of their classes. The total numbers of arrivals during consecutive time slots are i.i.d. random variables with arbitrary distribution. The classes of consecutively arriving customers, however, are correlated in a Markovian way, i.e., the probability that a customer belongs to a class depends on the class of the previously arrived customer. Service-time distributions are assumed to be general but class-dependent. We use probability generating functions to study the system analytically. The major aim of the paper is to estimate the impact of the *interclass correlation* in the arrival stream on the queueing performance of the system, in terms of the (average) number of customers in the system and the (average) customer delay and customer waiting time.

1 Introduction

Various types of scheduling disciplines have been investigated within the context of multi-class queueing systems. We mention, among others, priority scheduling (see, e.g., [4, 8, 11, 13, 15]), weighted fair queueing (WFQ) (see, e.g., [14, 17]), random order of service (ROS) (see, e.g., [1, 3, 10]), and generalized processor sharing (GPS) (see, e.g., [9,12,16]). Strangely enough, only few results have been derived for multi-class First-Come-First-Served (FCFS) systems, i.e., queueing systems in which the customers of different classes are accommodated in one queue and served in their order of arrival, irrespective of the classes they belong to (a recent paper is [5]). The present paper presents the analysis of a discrete-time model that fits in this category.

In classical multi-class queueing models, furthermore, it is generally assumed that the different classes occur randomly and independently in the arrival stream of customers (this is also the case in [5]). In this paper, however, we explicitly wish to examine the effect of so-called *interclass correlation* (or *class clustering*)

in the arrival process. Specifically, we are interested to know whether the degree to which customers of the same class have the tendency to arrive (and be served) closely together (i.e., back-to-back), or, conversely, the degree to which such customers have the tendency to be spread in time and mixed with customers of the other class, has a substantial impact on the performance of a *two-class* FCFS queueing system. In order to do so, we superimpose a two-state Markovian interclass correlation model (with arbitrary transition probabilities) on top of a regular general independent arrival process model for the aggregated customer stream. Service-time distributions are class-dependent but completely general. It is clear that the interclass correlation between consecutive customers can also be viewed as a form of non-independence between service times. One application of this queueing model is obvious: the two customers classes can model, for example, voice and data packets in a heterogeneous telecommunication system. It is common knowledge that data packets are significantly larger than voice packets. Then it is easy to see that if data packets have the tendency to arrive in clusters, the performance of the system may be degraded severely (with respect to voice packets). In this paper, we measure this degradation.

We first derive the probability generating function (pgf) of the total number of customers in the system at customer departure times. From this result, we can easily obtain the corresponding pgf valid at arbitrary slot boundaries. Various performance measures of practical use, such as the mean system content, the mean customer delay and the mean customer waiting time, can be easily derived from these pgf's by applying the moment-generating property of pgf's and by using Little's law. The resulting formulas and a number of numerical examples reveal that the system under study can exhibit two types of stochastic equilibrium, depending on the values of the system parameters: a "strong" equilibrium in which both customer classes individually generate less work than the system can handle (during periods where only such customers arrive), and a "compensated" type of equilibrium whereby one customer class creates overload situations which are compensated by strong under-load periods generated by the other customer class. In the latter case, our results clearly demonstrate the crucial importance of the amount of interclass correlation on the usual performance parameters of the system.

The outline of this paper is as follows. In Section 2, we describe the mathematical model. Section 3 first presents an analysis of the total number of customers in the system at customer departure times; next, the pgf of the system content at random slot boundaries is derived from this result. We discuss the results, both conceptually and quantitatively, in Section 4. Some conclusions are drawn in Section 5.

2 Mathematical model

We consider a discrete-time queueing system with infinite waiting room, one server, and two classes of customers, named A and B. As in all discrete-time models, the time axis is divided into fixed-length intervals referred to as *slots* in

the sequel. New customers may enter the system at any given (continuous) point on the time axis, but services are synchronized to (i.e., can only start and end at) slot boundaries. Customers are served in their order of arrival, regardless of the class they belong to. We call this service discipline "global FCFS".

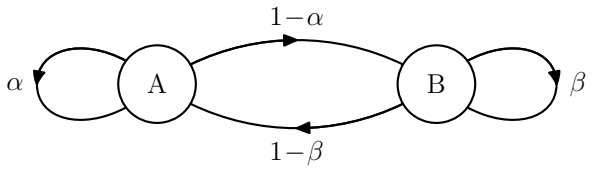

Fig. 1. Two-state Markov chain of the customer classes

The arrival process of new customers in the system is characterized in two steps. First, we model the total (aggregated) arrival stream of new customers by means of a sequence of i.i.d. non-negative discrete random variables (denoting the numbers of arrivals in consecutive slots) with common probability generating function (pgf) $E(z)$. The (total) mean number of arrivals per slot, in the sequel referred to as the (total) mean arrival rate, is given by $\lambda \triangleq E'(1)$. Next, we describe the occurrence of the two classes in the sequence of the consecutively arriving customers. In this study, we assume that both classes of customers account for part of the total load of the system, i.e., both customer classes are "mixed" in the arrival stream, but there may be some degree of "class clustering" in the arrival process, i.e., customers of any given class may (or may not) have a tendency to "arrive back-to-back". Mathematically, this means that the classes of two consecutive customers may be non-independent. Specifically, we assume a first-order Markovian type of correlation between the classes of two consecutively arriving customers, which basically means that the probability that the next customer belongs to a given class depends on the class of the previously arrived customer. Let t_k denote the class of customer k. The transition probabilities of the Markov chain that determines the class of the consecutively arriving customers are then defined as (see Fig. 1)

$$\text{Prob}[t_{k+1} = A \,|\, t_k = A] \triangleq \alpha \ , \tag{1}$$

$$\text{Prob}[t_{k+1} = B \,|\, t_k = A] \triangleq 1 - \alpha \ , \tag{2}$$

$$\text{Prob}[t_{k+1} = A \,|\, t_k = B] \triangleq 1 - \beta \ , \tag{3}$$

$$\text{Prob}[t_{k+1} = B \,|\, t_k = B] \triangleq \beta \ . \tag{4}$$

It is well known that for a two-state Markov chain of this type, the steady-state probabilities t_A and t_B of finding the chain in state A and B are given by

$$t_A \triangleq \lim_{k\to\infty} \text{Prob}[t_k = A] = \frac{1-\beta}{2-\alpha-\beta} \tag{5}$$

and

$$t_B \triangleq \lim_{k\to\infty} \text{Prob}[t_k = B] = \frac{1-\alpha}{2-\alpha-\beta} \ , \tag{6}$$

respectively (see, e.g., [2]). The quantities t_A and t_B can be interpreted as the fractions of class A and class B customers in the arrival stream. The (steady-state) correlation coefficient of the Markov chain, i.e., the amount of correlation between the classes of two consecutively arriving customers (in the steady state), is given by

$$\gamma \triangleq \lim_{k\to\infty} \frac{\text{E}[t_k t_{k+1}] - \text{E}[t_k]\text{E}[t_{k+1}]}{\sqrt{\text{var}[t_k]\,\text{var}[t_{k+1}]}} = \alpha + \beta - 1 \ . \tag{7}$$

We will indicate the parameter γ ($-1 \leq \gamma \leq +1$) as the *interclass correlation* in the sequel. Positive values of γ correspond to a situation whereby the customers of any given class have a tendency to cluster, while negative values of γ refer to arrival streams in which the customers of classes A and B have a tendency to alternate, i.e., be mixed more strongly. The case where $\gamma = 0$, of course, corresponds to the classical assumption that classes of subsequent customers are independent.

The service process of the system is characterized by attaching to each customer a corresponding *service time*, which indicates the number of time slots required to give complete service to the customer at hand. The service times of customers are class-dependent and are modelled as a sequence of independent positive discrete random variables with pgf's $A(z)$ and $B(z)$. The corresponding mean values are given by $\mu_A \triangleq A'(1)$ and $\mu_B \triangleq B'(1)$ for customers of class A and B, respectively.

3 System analysis

3.1 System equations at customer departure times

Let u_k denote the total *system content*, i.e., the total number of customers present in the system just after the service completion of the k-th customer, and, as before, let t_k indicate the class customer k belongs to. Then, as a consequence of all the model assumptions in Section 2, the couple (t_k, u_k) forms a Markovian state description of the system (at customer departure times).

The state transitions of the quantities $\{t_k\}$ are governed by the Eqs. (1)-(4), whereas for the quantities $\{u_k\}$, the following recursive system equations can be established (see Figs. 2 and 3):

$$u_{k+1} = \begin{cases} u_k - 1 + g_{k+1} & \text{if } u_k > 0 \\ f_{k+1} + g_{k+1} & \text{if } u_k = 0 \end{cases} . \tag{8}$$

Here, the quantity g_{k+1} is defined as the number of arrivals in the system during the service time of customer $k+1$, while f_{k+1} indicates the number of customers arriving *after* customer $k+1$ in its arrival slot.

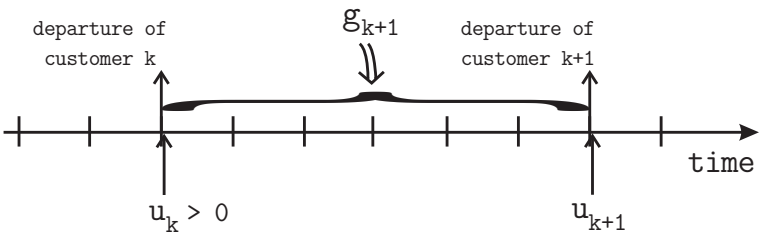

Fig. 2. Relationship between u_k and u_{k+1} when $u_k > 0$

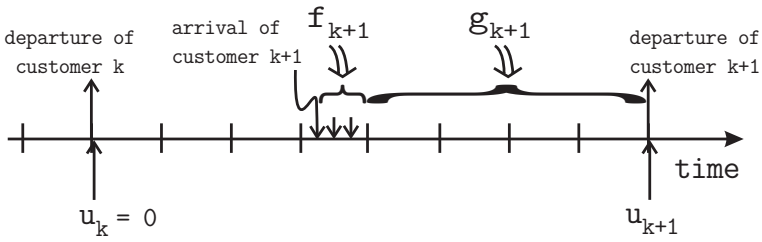

Fig. 3. Relationship between u_k and u_{k+1} when $u_k = 0$

It is easily seen that the pgf of f_{k+1} is given by the pgf of the number of additional arrivals in a slot with at least one arrival, i.e.,

$$F(z) \triangleq \mathrm{E}[z^{f_{k+1}}] = \frac{E(z) - E(0)}{z[1 - E(0)]} \ , \tag{9}$$

regardless of the class of customer $k+1$. The distribution of the quantity g_{k+1}, however, does depend on the class of customer $k+1$. We have

$$G_A(z) \triangleq \mathrm{E}[z^{g_{k+1}} \,|\, t_{k+1} = A] = A(E(z)) \ , \tag{10}$$

$$G_B(z) \triangleq \mathrm{E}[z^{g_{k+1}} \,|\, t_{k+1} = B] = B(E(z)) \ . \tag{11}$$

3.2 System content at customer departure times

Let us assume that the queueing system at hand is stable, i.e., that the stability condition is fulfilled. Intuitively, it is not difficult to see that the system is stable if and only if the average amount of work entering the system per slot is strictly less than 1, i.e., if and only if $\lambda \mathrm{E}[c] < 1$, with $\mathrm{E}[c]$ the average service time of an arbitrary customer. Expressed in the basic parameters of our system, this is equivalent to the condition

$$\lambda(t_A \mu_A + t_B \mu_B) < 1 \ , \tag{12}$$

where the quantities t_A and t_B are the steady-state probabilities of the arrival Markov chain (see Eqs. (5) and (6)). Assuming this condition fulfilled, we define

the joint steady-state probabilities of the Markov chain $\{(t_k, u_k)\}$ as

$$p_A(i) \triangleq \lim_{k\to\infty} \text{Prob}[t_k = A, u_k = i] \tag{13}$$

and

$$p_B(i) \triangleq \lim_{k\to\infty} \text{Prob}[t_k = B, u_k = i] \ , \tag{14}$$

for all $i \geq 0$. The corresponding partial pgf's are defined as $P_A(z)$ and $P_B(z)$. Then the steady-state pgf $P(z)$ of the total system content at customer departure times is equal to $P_A(z) + P_B(z)$.

We now establish two linear equations for $P_A(z)$ and $P_B(z)$. We depart from the balance equations of the Markov chain $\{(t_k, u_k)\}$ for class A:

$$\begin{aligned} p_A(j) = &\sum_{i=0}^{\infty} p_A(i)\alpha \lim_{k\to\infty} \text{Prob}[u_{k+1} = j \,|\, t_{k+1} = A, u_k = i] \\ &+ \sum_{i=0}^{\infty} p_B(i)(1-\beta) \lim_{k\to\infty} \text{Prob}[u_{k+1} = j \,|\, t_{k+1} = A, u_k = i] \ . \end{aligned} \tag{15}$$

Next, we introduce pgf's into this equation:

$$\begin{aligned} P_A(z) = &\alpha \sum_{i=0}^{\infty} p_A(i) \lim_{k\to\infty} \text{E}[z^{u_{k+1}} \,|\, t_{k+1} = A, u_k = i] \\ &+ (1-\beta) \sum_{i=0}^{\infty} p_B(i) \lim_{k\to\infty} \text{E}[z^{u_{k+1}} \,|\, t_{k+1} = A, u_k = i] \ . \end{aligned} \tag{16}$$

The expected values in (16) can be further developed by using the system equations (see Eq. 8):

$$\begin{aligned} \lim_{k\to\infty} \text{E}[z^{u_{k+1}} \,|\, t_{k+1} = A, u_k = i] &= \lim_{k\to\infty} \text{E}[z^{i-1+g_{k+1}} \,|\, t_{k+1} = A] \\ &= z^{i-1} G_A(z) \ , \end{aligned} \tag{17}$$

for $i \geq 1$, and

$$\begin{aligned} \lim_{k\to\infty} \text{E}[z^{u_{k+1}} \,|\, t_{k+1} = A, u_k = 0] &= \lim_{k\to\infty} \text{E}[z^{f_{k+1}+g_{k+1}} \,|\, t_{k+1} = A] \\ &= F(z) G_A(z) \ . \end{aligned} \tag{18}$$

Putting everything together, we then obtain

$$\begin{aligned} P_A(z) = &\alpha \frac{G_A(z)}{z} [P_A(z) - P_A(0)] + \alpha P_A(0) F(z) G_A(z) \\ &+ (1-\beta) \frac{G_A(z)}{z} [P_B(z) - P_B(0)] + (1-\beta) P_B(0) F(z) G_A(z) \ . \end{aligned} \tag{19}$$

Using Eqs. (9)-(11), we finally obtain a first linear equation between $P_A(z)$ and $P_B(z)$:

$$\begin{aligned}[z - \alpha A(E(z))]\, P_A(z) - (1-\beta)A(E(z))P_B(z) \\ = \frac{E(z)-1}{1-E(0)}\left[\alpha P_A(0) + (1-\beta)P_B(0)\right] A(E(z)) \ .\end{aligned} \tag{20}$$

Starting from the balance equations for class B, we can derive a second, similar equation:

$$\begin{aligned}[z - \beta B(E(z))]\, P_B(z) - (1-\alpha)B(E(z))P_A(z) \\ = \frac{E(z)-1}{1-E(0)}\left[\beta P_B(0) + (1-\alpha)P_A(0)\right] B(E(z)) \ .\end{aligned} \tag{21}$$

Eqs. (20) and (21) form a set of two linear equations for the two unknown partial pgf's $P_A(z)$ and $P_B(z)$. Expressions for these pgf's can be found by solving the set. Then adding up $P_A(z)$ and $P_B(z)$ leads to the following expression for the pgf $P(z)$:

$$\begin{aligned}P(z) = &\frac{P(0)(E(z)-1)}{1-E(0)} \\ &\times \frac{z[p_A A(E(z)) + p_B B(E(z))] + (1-\alpha-\beta)A(E(z))B(E(z))}{z^2 - z[\alpha A(E(z)) + \beta B(E(z))] - (1-\alpha-\beta)A(E(z))B(E(z))} \ ,\end{aligned} \tag{22}$$

where the quantities p_A and p_B are defined as

$$p_A \triangleq \frac{\alpha P_A(0) + (1-\beta)P_B(0)}{P(0)} \tag{23}$$

and

$$p_B \triangleq \frac{(1-\alpha)P_A(0) + \beta P_B(0)}{P(0)} \ , \tag{24}$$

respectively. It is not difficult to see that p_A and p_B denote the conditional probabilities that a customer entering an empty system (in the steady state) belongs to class A or B: $p_X \triangleq \lim_{k\to\infty} \mathrm{Prob}[t_{k+1} = X \,|\, u_k = 0]$, with $X \in \{A, B\}$.

The probability $P(0)$ can be derived explicitly from the normalization condition of the pgf $P(z)$, i.e., the condition $P(1) = 1$. The result is given by

$$P(0) = \frac{1-E(0)}{\lambda}\ [1 - \lambda(t_A\mu_A + t_B\mu_B)] = \frac{1-E(0)}{\lambda}\ \{1 - \lambda \mathrm{E}[c]\} \ , \tag{25}$$

where, as before, the quantities t_A and t_B are the steady-state probabilities of the arrival Markov chain, defined in Eqs. (5) and (6), and $\mathrm{E}[c]$ denotes the average service time of an arbitrary customer. It then remains for us to calculate the two unknown probabilities p_A and p_B, of which we know from (23) and (24) that $p_A + p_B = 1$. The unknowns can be determined, in general, by invoking the

well-known property that pgf's such as $P(z)$ are bounded inside the closed unit disk $\{z : |z| \leq 1\}$ of the complex z-plane, at least when the stability condition (12) of the queueing system is met (only in such a case our analysis was justified and $P(z)$ can be viewed as a legitimate pgf). Now, it can be shown by means of Rouché's theorem from complex analysis [2,7] that the denominator of Eq. (22) has exactly two zeroes inside the closed unit disk of the complex z-plane, one of which is equal to 1, as soon as the stability condition (12) is fulfilled. It is clear that these two zeroes should also be zeroes of the numerator of Eq. (22), as $P(z)$ must remain bounded in those points. For the zero $z = 1$, this condition is fulfilled regardless of the values of the unknowns p_A and p_B, since the numerator of (22) contains a factor $E(z) - 1$. However, for the second zero, say $z = \hat{z}$, the requirement that the numerator should vanish yields a linear equation for the two unknowns. A second linear equation is given by $p_A + p_B = 1$. In general, i.e., when the pgf's $A(z)$ and $B(z)$ are different, the two unknown probabilities p_A and p_B can be found as the solutions of the two established linear equations. We obtain

$$p_A = \frac{\alpha A(E(\hat{z})) - (1-\beta)B(E(\hat{z})) - \hat{z}}{A(E(\hat{z})) - B(E(\hat{z}))} \tag{26}$$

and

$$p_B = \frac{\beta B(E(\hat{z})) - (1-\alpha)A(E(\hat{z})) - \hat{z}}{B(E(\hat{z})) - A(E(\hat{z}))} \quad . \tag{27}$$

Once the zero $\hat{z}$ has been computed (numerically), p_A and p_B can be derived from (26) and (27). Substitution of the obtained values and of Eq. (25) in (22) then leads to a fully determined expression of the steady-state pgf $P(z)$ of the total system content at customer departure times.

3.3 System content at random slot boundaries

It has been shown in [2] that in any discrete-time queueing system with one single server and independent arrivals from slot to slot (with pgf $E(z)$), regardless of the precise characteristics of the service process and the intra-slot details of the arrival process (the position of the arrival instants within the slot, single arrivals or batch arrivals, etc.), the following simple relationship is valid between the pgf $S(z)$ of the system content at random slot boundaries and the pgf $P(z)$ valid at customer departure times:

$$P(z) = \frac{E(z) - 1}{\lambda(z-1)} \; S(z) \quad . \tag{28}$$

In the previous subsection, we have found an expression for the pgf $P(z)$. Hence, it is easy to obtain an expression for $S(z)$. From $S(z)$, various performance measures of practical importance can be derived. For instance, the mean system

content at random slot marks can be found as $\mathrm{E}[s] = S'(1)$. After long and tedious calculations, this results in

$$\mathrm{E}[s] = \rho + \frac{\lambda^2 C''(1) + E''(1)C'(1)}{2(1-\rho)} + \frac{\gamma t_A t_B \lambda^2 (\mu_A - \mu_B)^2}{(1-\gamma)(1-\rho)} + \frac{\lambda(p_A - t_A)(\mu_A - \mu_B)}{1-\gamma}, \tag{29}$$

where t_A and t_B are expressed in Eqs. (5) and (6), γ is the interclass correlation defined in (7), $C'(1)$ and $C''(1)$ are derivatives of the pgf $C(z)$ of the service time of an arbitrary customer (i.e., $C(z) \triangleq t_A A(z) + t_B B(z)$), ρ $(= \lambda C'(1))$ is the total load of the queueing system, and p_A and p_B are the unknown probabilities defined in (23) and (24) and calculated as (26) and (27) (as soon as the zero $\hat{z}$ has been determined numerically). The first term (ρ) in Eq. (29) corresponds to the mean number of customers in service, the other three terms account for the mean *queue content*, i.e., the mean number of customers that are actually waiting to be served.

Higher-order moments of the system-content distribution can be obtained by computing higher-order derivatives of the pgf $S(z)$. By applying (the discrete-time version of) Little's law (see, e.g., [6]), the mean *delay* (system time) of an arbitrary customer can be obtained as $\mathrm{E}[d] = \mathrm{E}[s]/\lambda$. The mean *waiting time* of an arbitrary customer can be derived from this as $\mathrm{E}[w] = \mathrm{E}[d] - \mathrm{E}[c]$:

$$\mathrm{E}[w] = \frac{\lambda^2 C''(1) + E''(1)C'(1)}{2\lambda(1-\rho)} + \frac{\gamma t_A t_B \lambda (\mu_A - \mu_B)^2}{(1-\gamma)(1-\rho)} + \frac{(p_A - t_A)(\mu_A - \mu_B)}{1-\gamma}. \tag{30}$$

4 Discussion of results and numerical examples

In this section, we discuss the results, both from a qualitative perspective and by means of some numerical examples. The first interesting result obtained is the form of the stability condition $\lambda < \dfrac{1}{\mathrm{E}[c]} = \dfrac{1}{t_A \mu_A + t_B \mu_B}$, which shows that the maximum achievable throughput of this system, expressed in customers per slot, is completely determined by the mean service time of an arbitrary customer, regardless of the possible interclass correlation.

Next, we focus on the mean system content at random slot marks (see Eq. (29)). This result explicitly and very clearly shows the influence of the various system parameters on the performance of the system. As could be expected intuitively, the mean system content depends on the first two moments of the arrival process (as represented by the quantities λ and $E''(1)$, and to some extent $\rho = \lambda C'(1)$) and the first two moments of the service times (contained in the quantities $C'(1)$, $C''(1)$, μ_A, μ_B, and also $\rho = \lambda C'(1)$). It is not surprising that $\mathrm{E}[s]$ goes to infinity as ρ approaches its limiting value 1, dictated by the stability condition of the system. However, it is striking that $\mathrm{E}[s]$ also seems to increase without bound if the interclass correlation $\gamma = \alpha + \beta - 1$ approaches

the value +1, even when the stability condition $\rho < 1$ is met. Positive interclass correlation appears to be very detrimental for the performance of the system, whereas negative interclass correlation has a very moderate positive effect on the performance.

The first two terms in Eq. (29) correspond to the classical result that would be obtained in a system without interclass correlation and with service-time pgf $C(z)$ (see, e.g., [2]). This means that the third and fourth term in (29) can be fully attributed to the presence of interclass correlation in the arrival process. We note, indeed, that the third term vanishes when $\gamma = 0$; in the fourth term, both t_A and p_A reduce to the same value α when $\gamma = 0$ (see Eqs. (5) and (26), with $\hat{z} = 0$), which implies that the fourth term is equal to zero as well in that case. It is easy to see that the third and fourth term also disappear when all customers have the same service-time distribution, i.e., when $A(z) = B(z)$ and, hence, $\mu_A = \mu_B$, and, finally, when there is only one class of customers in the system, i.e., when either $\alpha = 1$ (and, hence, $p_A = t_A = 1$ and $t_B = 0$) or $\beta = 1$ (and, therefore, $p_A = t_A = 0$).

Let us now consider some numerical results. In a first example, we assume Poisson arrivals (i.e., $E(z) = e^{\lambda(z-1)}$), equal fractions of both classes of customers in the arrival stream (i.e., $t_A = t_B = 0.5$), geometrically distributed service times for both classes, i.e.,

$$X(z) = \frac{z}{\mu_X + (1 - \mu_X)z} \quad , \tag{31}$$

with $X \in \{A, B\}$, and with $\mu_A = 8$ and $\mu_B = 2$. The stability condition is then given by $\rho = \lambda[t_A\mu_A + t_B\mu_B] = 5\lambda < 1$ (i.e., $\lambda < 0.2$).

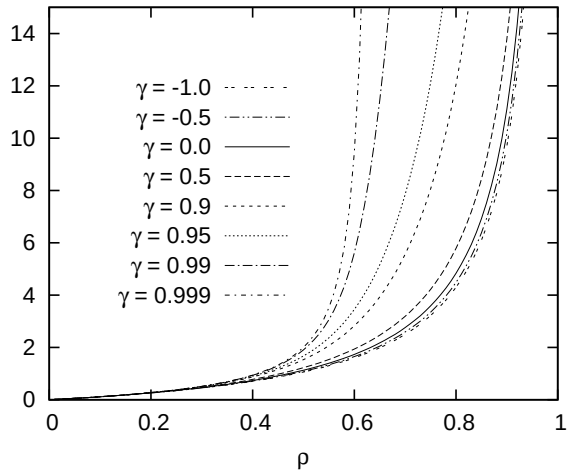

Fig. 4. Mean system content E[s] versus load ρ for various values of the interclass correlation γ

Fig. 4 shows the mean system content E[s] as a function of the load ρ, for various values of the interclass correlation γ. The figure confirms that, for given values of $\rho < 1$, the parameter γ has a major impact on the results when it is positive and only a minor influence when it is negative. An intuitive explanation of this phenomenon lies in the observation that the numbers of consecutive class-A and class-B customers in the arrival stream both increase dramatically as γ approaches the value $+1$. Indeed, the mean number of class-A (class-B) customers between two consecutive class-B (class-A) customers is given by $\frac{1}{1-\alpha} = \frac{2}{1-\gamma}$ ($\frac{1}{1-\beta} = \frac{2}{1-\gamma}$). For negative values of γ, this implies that customers of both classes alternate strongly; for positive values of γ, there may be very long sequences of customers of the same class. During such periods, the momentary load is either given by $\rho_A \triangleq \lambda\mu_A = 8\lambda$ or by $\rho_B \triangleq \lambda\mu_B = 2\lambda$. It is easily seen that the stability condition $\rho < 1$, or $\lambda < 0.2$, guarantees that $\rho_B < 1$, but not necessarily that $\rho_A < 1$. It is clear that if λ or ρ are small enough (more specifically, $\lambda < 0.125$ or $\rho < 0.625$), $\rho_A < 1$ and $\rho_B < 1$, i.e., the system is locally stable both during A- and B-sequences (and, hence, also globally stable - we call this the "strong" equilibrium), while if $0.125 \leq \lambda < 0.2$, or, equivalently, $0.625 \leq \rho < 1$, $\rho_B < 1$ but $\rho_A > 1$, i.e., the system is locally stable during B-sequences but not during A-sequences. In the latter case, labelled as the "compensated" equilibrium, (global) stability is assured because although the queue size builds up during A-sequences (because, on average, more work arrives than the server can perform), it decreases again during B-sequences (when much less work enters than the server can execute). In other words, the overload periods created by the A-customers are *compensated* by the underload periods of the B-customers. When the interclass correlation approaches $+1$, however, the amplitude of these queue size variations goes to infinity, implying that the mean system content does the same.

The same behavior can be observed in Fig. 5, where we have plotted E[s] as a function of γ for various values of ρ. The figure illustrates very clearly that the system content grows without bound as $\gamma \to +1$ when ρ is higher than its critical value 0.625. When ρ is less than this critical value, on the other hand, the mean system content remains finite for all values of γ. Although we have explained this behaviour intuitively in the previous paragraph, it is somewhat unexpected in view of Eq. (29). Indeed, Eq. (29) seems to say that E[s] should become unbounded as $\gamma \to +1$, *regardless of the other system parameters*. The third and fourth term in (29) both approach infinity for $\gamma \to 1$; however, when ρ is less than its critical value, the terms cancel each other causing their sum to remain finite.

A second example is treated in Figs. 6 and 7. Again, we assume Poisson arrivals and geometrically distributed service times for both classes. Here, however, $\mu_A = 100$ and $\mu_B = 10$. The interclass correlation γ is kept constant at 0.8. This implies that $\alpha = 0.8 + 0.2t_A$, $\beta = 1 - 0.2t_A$, and $\rho = 10\lambda(1 + 9t_A)$. We now investigate the impact of the parameter t_A, i.e., the fraction of class-A customers in the arrival stream, on the mean system content and the mean waiting times of

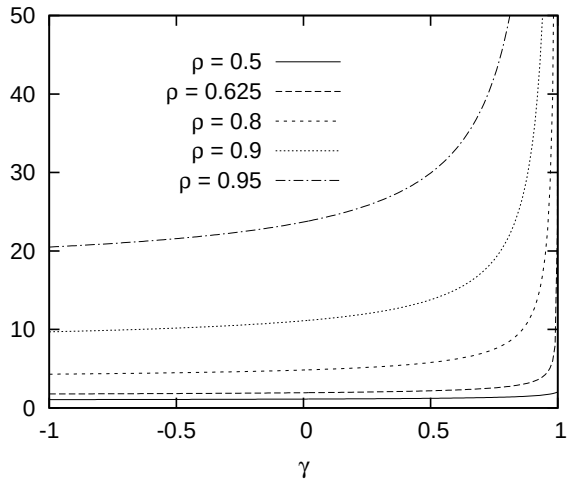

Fig. 5. Mean system content E[s] versus interclass correlation γ for various values of the load ρ

the customers. Fig. 6 shows the mean system content E[s] versus t_A, for various values of ρ, whereas Fig. 7 illustrates the corresponding results for the mean waiting time E[w] of the customers. Fig. 6 reveals that, for any given value of the total load ρ, the mean system content increases as a function of t_A for "low" values of t_A (more or less in the interval $0 \leq t_A \leq 0.1$), then reaches a maximum value for t_A somewhere around 0.1, and, finally decreases monotonically in the interval $0.1 \leq t_A \leq 1$. An intuitive explanation might be as follows. For $t_A = 0$, all customers belong to class B (with a short service time of 10 slots); as soon as t_A becomes positive, say $0 \leq t_A \leq 0.1$, class-A customers (with a long service time of 100 slots) arrive sporadically and (when in service) somehow block the regular processing of class-B customers, which causes the system content to increase. If, however, t_A increases further (while the total load ρ remains constant), the system receives considerably less customers (for the same amount of work), which explains the decreasing system content in the interval $0.1 \leq t_A \leq 1$.

The behaviour of the mean waiting time (see Fig. 7) is qualitatively a bit similar as for the mean system content. More specifically, it can be observed that the mean waiting times also increase for "low" values of t_A to reach a maximum value and then decrease for "higher" values of t_A. However, the maximum value of the waiting time is attained for t_A around 0.25, whereas the highest mean system content occurs for t_A in the vicinity of 0.1. Also, the rates at which the mean waiting times increase and decrease seem relatively slower than for the mean system content. Intuitively, this can be attributed to the fact that the waiting time reflects the unfinished work in the system (at the arrival instant of a customer), while the system content indicates the number of customers in the system, whereby all customers contribute identically, irrespective of their service

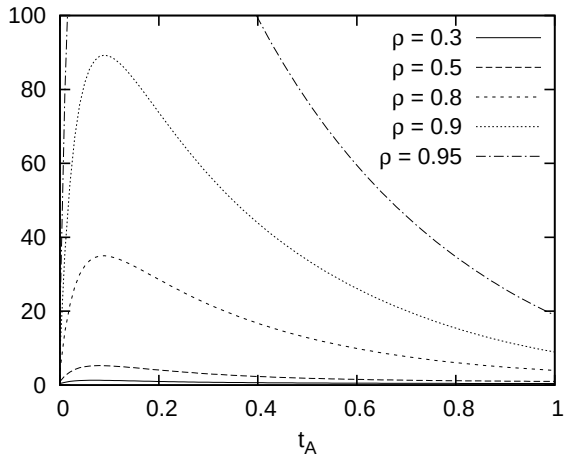

Fig. 6. Mean system content E[s] versus the fraction t_A of class-A customers for various values of the load ρ

time, i.e., irrespective of the amount of work they represent. The fact that the mean waiting time (and, hence, the unfinished work in the system) for $t_A = 0$ is substantially smaller than for all other values of t_A can be explained by the higher burstiness of the arrival process of work units if class-B customers (bringing small amounts of work) are alternated with class-A customers (bringing large batches of work at the same time), which happens as soon as t_A gets positive.

5 Conclusions

In this paper, we have studied a discrete-time queueing system with one server and two classes of customers, and operating under the global FCFS service discipline. We have assumed independent (aggregated) arrivals from slot to slot combined with a general first-order Markovian interclass correlation model, and general but class-dependent service-time distributions. We have been able to derive the main performance measures of the system in semi-analytical form, i.e., we have obtained explicit expressions for such quantities as the mean system content and the mean customer waiting time in terms of the basic parameters of the model and one parameter which is only implicitly known through a non-linear equation that it satisfies.

The results reveal that the interclass correlation does not have any effect on the stability condition of the system, but it may have a very direct and great influence on the main performance measures of the system. More specifically, when the system is (globally) stable, we have observed that two different kinds of global equilibrium are possible, depending on the exact value of the load. For "low" values of the load, the system exhibits a "strong" equilibrium, whereas

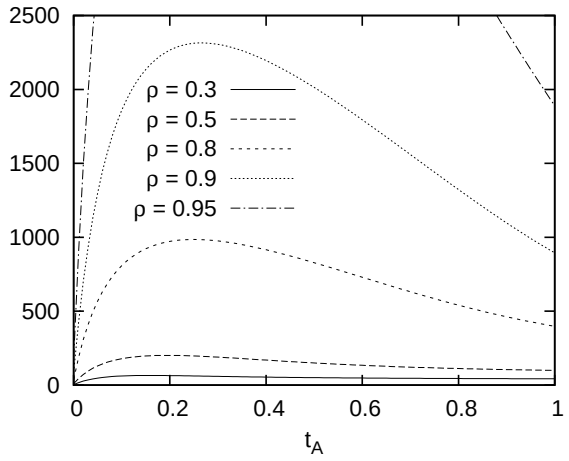

Fig. 7. Mean waiting time E[w] versus the fraction t_A of class-A customers for various values of the load ρ

for higher loads, the system reaches a "compensated" type of equilibrium. Especially in the latter case, the impact of strong positive interclass correlation may be devastating for the queueing performance. We therefore believe that the phenomenon of class clustering in the context of multi-class queueing systems deserves more attention than it traditionally has received in the classical queueing literature.

Acknowledgment

The last two authors are Postdoctoral Fellows with the Fund for Scientific Research, Flanders (F.W.O.-Vlaanderen), Belgium.

References

1. S.C. Borst, O.J. Boxma, J.A. Morrison, and R.N. Queija. The equivalence between processor sharing and service in random order. *Operations Research Letters*, 31(4):254–262, 2003.
2. H. Bruneel and B.G. Kim. *Discrete-time models for communication systems including ATM*. Kluwer Academic, Boston, USA, 1993.
3. G.M. Carter and R.B. Cooper. Queues with service in random order. *Operations Research*, 20(2):389–405, 1970.
4. J. Chen and R. Guérin. Performance study of an input queueing packet switch with two priority classes. *IEEE Transactions on Communications*, 39(1):117–126, 1991.
5. S. De Clercq, K. Laevens, B. Steyaert, and H. Bruneel. A multi-class discrete-time queueing system under the fcfs service discipline. *Annals of Operations Research*. Accepted for publication.

6. D. Fiems and H. Bruneel. A note on the discretization of Little's result. *Operations Research Letters*, 30:17–18, 2002.
7. M.O. Gonzáles. *Classical complex analysis*. Marcel Dekker, New York, USA, 1992.
8. N. Jaiswal. *Priority queues*. Academic Press, New York, 1968.
9. X. Jin and G. Min. Analytical modelling and evaluation of generalized processor sharing systems with heterogeneous traffic. *International Journal of Communication Systems*, 21(6):571–586, 2008.
10. J. Kim, J. Kim, and B. Kim. Analysis of the M/G/1 queue with discriminatory random order service policy. *Performance Evaluation*, 68(3):256–270, 2011.
11. K. Laevens and H. Bruneel. Discrete-time multiserver queues with priorities. *Performance Evaluation*, 33(4):249–275, 1998.
12. P. Lieshout and M. Mandjes. Generalized processor sharing: Characterization of the admissible region and selection of optimal weights. *Computers & Operations Research*, 35(8):2497–2519, 2008.
13. T. Maertens, J. Walraevens, and H. Bruneel. Performance comparison of several priority schemes with priority jumps. *Annals of Operations Research*, 180(3):1168–1185, 2008.
14. J.F. Shortle and M.J. Fischer. Approximation for a two-class weighted fair queueing discipline. *Performance Evaluation*, 67(10):946–958, 2010.
15. J. Walraevens, D. Fiems, S. Wittevrongel, and H. Bruneel. Calculation of output characteristics of a priority queue through a busy period analysis. *European Journal of Operational Research*, 198(3):891–898, 2009.
16. J. Walraevens, J.S.H. van Leeuwaarden, and O.J. Boxma. Power series approximations for two-class generalized processor sharing systems. *Queueing systems*, 66(2):107–130, 2010.
17. L. Wang, G. Min, D.D. Kouvatsos, and X. Jin. Analytical modeling of an integrated priority and WFQ scheduling scheme in multi-service networks. *Computer Communications*, 33:S93–S101, 2010.